Affordance Diffusion: Synthesizing Hand-Object Interactions Supplementary Materials

Yufei Ye[1*] Xueting Li[2] Abhinav Gupta[1] Shalini De Mello[2]
Stan Birchfield[2] Jiaming Song[2] Shubham Tulsiani[1] Sifei Liu[2]
[1]Carnegie Mellon University [2]NVIDIA
https://judyye.github.io/affordiffusion-www

In the supplementary material, we provide more implementation details and more qualitative results. We discuss the details of articulation-agnostic hand proxy and how to apply DDPM loss in the image space for training the LayoutNet (Sec. A.1). We also present ablations on ContentNet(Sec. A.2). We further show: (i) the paired data construction method being robust, in Sec. A.3, (ii) baseline implementations details in Sec. A.4, (iii) details of integrating our approach to scene-level affordance prediction in Sec. A.5. Finally, we discuss the limitation of our approach (Sec. A.6), and show more qualitative results in Sec. B. **Visual results are also included in the video.**

A. Implementation Details

A.1. LayoutNet (Sec 3.1)

Layout parameters. As mentioned in Sec 3.1 of the main paper, we parameterize the layout as (x, y, a, b_1, b_2), where x, y is the location, a^2 is size, and b_1, b_2 are un-normalized approaching direction parameters. For training the LayoutNet, we obtain the ground truth parameters from off-the-shelf 2D hand prediction systems. The size and location comes from the predicted bounding box of a hand detector [10], which typically defines the hand region up to the wrist. The orientation is calculated from hand segmentation whose region is typically defined as the entire hand region, including hand and forearm. The approaching direction is calculated as the first principal component of a hand mask that centers on the location of the palm of the predicted hand.

We splat the layout parameters onto 2D via the spatial transformer network [4] that transforms a canonical mask template by a similarity transformation. The 2D similarity transformation is determined from the layout parameters. More formally,

$$T_l = \begin{pmatrix} sR & t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a^2\hat{b}_1 & -a^2\hat{b}_2 & x \\ a^2\hat{b}_2 & a^2\hat{b}_1 & y \\ 0 & 0 & 1 \end{pmatrix},$$

where $\hat{b}_1, \hat{b}_2$ is the normalized vector of b_1, b_2.

The lollipop-shape template in the canonical space is implemented with its circle being an isometric 2D Gaussian with a standard deviation of 1 and its rectangle being a 1D Gaussian with a standard deviation $\bar{s} = 2$. The width of the rectangle is calculated from the training data as the average ratio of the widths of forearms and palms.

DDPM loss on mask. In Eq 1 and 2 of the main paper, we write the DDPM loss in terms of reconstructing clean samples. In practice, we follow prior works [6–8] that reconstruct the added noise ϵ as

$$\mathcal{L}^{\text{noise}}_{\text{DDPM}} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,I),t}\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2.$$

The estimated clean sample $\hat{l}_0$ is connected with the estimated noise by $\hat{l}_0 = \frac{1}{\sqrt{1-\bar{\alpha}_t}} l_t - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta$, where $\alpha_t, \bar{\alpha}_t$ represent the noise schedule for each diffusion time step.

We train the LayoutNet with a weighted sum of the parameter loss $\mathcal{L}_{\text{para}}$ for esitmating the noise term ϵ, and a mask loss $\mathcal{L}_{\text{mask}}$ for estimating the clean sample term $\hat{l}_0$. The hyperparamter λ is set to 10.

Guided layout generation. LayoutNet inherits properties from diffusion models that can be guided to generate samples with additional constraints at test time. We follow Song *et al.* [11]. After each diffusion steps, we hijack the additional constraints with corresponding noise levels for the next diffusion step.

More specifically, instead of passing in the network's output $\boldsymbol{x}_t$ from the previous time step, we hijack it with $\boldsymbol{x}_t \leftarrow \tilde{\boldsymbol{x}}_t\boldsymbol{m} + \boldsymbol{x}_t(1-\boldsymbol{m})$, where $\boldsymbol{m}$ is the indicator mask of the given condition $\tilde{\boldsymbol{x}}_0$. The unspecified constraints in $\tilde{\boldsymbol{x}}_0$ are set to 0. $\tilde{\boldsymbol{x}}_t$ represents the additional constraint with corresponding noise level, *i.e.* $\sqrt{1-\bar{\alpha}_t}\tilde{\boldsymbol{x}}_0 + \sqrt{\bar{\alpha}_t}\boldsymbol{\epsilon}$.

*Yufei was an intern at NVIDIA during the project.

A.2. ContentNet (Sec3.2)

The goal of ContentNet is to generate high-resolution (256^2) realistic HOI images conditioned on the predicted layout and the input object image. We tried two different approaches commonly used in diffusion models [6, 8] as backbones for the ContentNet. One way (called ours/AffordDiff-LDM) is to follow Rombach *et al*. [9], as described in our main paper, that implements the ContentNet in the latent space where images of size 256^2 are compressed to 3-dimensional features of size 64^2 by a fixed pretrained autoencoder. The other way (called ours/AffordDiff-GLIDE) is to follow Nichol *et al*. [6] that uses a cascaded diffusion model that first generates images of size 64^2 and then upsamples them by a factor of 4.

All of the quantitative results in our main paper, including the user studies and all ablations, are based on Afford-LDM. AffordDiff-GLIDE is better in terms of contact recall (90.8% vs 87.1%) while AffordDiff-LDM is significantly better in terms of FID score (99.0 vs 121.6). We find that AffordDiff-LDM generates less blurry results and the hand texture appears sharper and more realistic. In comparison, we find AffordDiff-GLIDE perceptually preferred because AffordDiff-GLIDE generates more realistic, though blurrier, finger articulations. The qualitative results in the main paper on EPIC-KITCHEN dataset (Fig 1 and Fig4 right in the main paper) show Afford-GLIDE. However, we provide the qualitative comparison of Afford-LDM with baselines in Fig 1 and Fig 2 of the appendix. We further provide a comparison of these two variants in Fig 7 of the appendix.

A.3. Constructing Paired Training Data (Sec3.3)

Cropping Details. We crop all objects with 80% squared padding before resizing such that objects (hands) appear in similar (different) sizes. The model learns the priors of their relative scales, *e.g*., a hand to grasp a kettle appears much smaller than that of a mug (Fig 4).

We show that the proposed method to obtain pixel-aligned pairs of HOI and object-only images is robust and can also be applied to more cluttered images. When there is more than one hand in the HOI image, we randomly select one to remove. We show results of applying our data construction method on the HOI4D (Fig 1) and the EPIC-KITCHEN (Fig 2) datasets.

A.4. Baselines Implementation

Pix2Pix [3] (Sec4.1) We modify the official Pix2Pix implementation[1]. Given the predicted layout and the provided object image, we concatenate them channel-wise and pass them through 6 blocks of ResNet to output HOI images. The discriminator takes in the concatenation the of the object-only image, the splatted layout image, and generated HOI image and learns to discriminate between the real and fake domains. We tried batchnorm and instancenorm and found that batchnorm generated better results in general but has some black holes if the background statistics deviate from that of the training set.

VAE [5] (Sec4.1) VAE is notoriously known for being hard to balance for both generation variance and reconstruction quality. We sweep hyperparameters of the KL divergence loss's weights from $1, 1e-1, 1e-2, 1e-3, 1e-4$ and use $1e-3$ as it produces the highest contact recall.

GANHand [2] (Sec4.2) GANHand is originally proposed both to predict 3D MANO hands for images of YCB objects [1] and to optimize physical plausibility with respect to the known or reconstructed 3D shapes of YCB objects. We compare our method with their sub-network for grasp prediction from RGB images (blue branch in their original paper, Fig 4). The sub-network takes in the object's identity, the desk plane equation and the object's center in 3D space, in addition to the object image. Since these are not available in the HOI4D dataset, we set them to zeros. We apply an additional reconstruction loss for 3D hand joints, MANO hand parameters and camera parameters. We finetune the network from the public checkpoints for another 10k iterations.

A.5. Scene Integration

We integrate our object-centric HOI synthesis to scene-level affordance prediction. We first detect the objects in the scene and then expand the detected bounding box's size with the same pad ratio (0.8 of the original object size). However, when the scene is crowded, the extended object crops may include other objects thus distracting the layout generation. We instead crop the object with the detected bounding box and pad the cropped object with boundary values. This allows the network to generate hand interaction only for the object of interest.

A.6. Limitation and Failure Cases

Although it is encouraging that the proposed model can perform zero-shot generalization to the EPIC-KITCHEN dataset, the proposed method inherits limited generalization capabilities from general learning-based algorithms. The proposed model will fail when the object image's appearance deviates too much from the training set, *e.g.* for too cluttered scenes, extreme lighting, very large objects (like a fridge) or very small objects (like a pin), *etc*. The current model also cannot generate hands entering from the top of the frame or generate hands from a third-person's view due to the bias in the training set. These limitations require training with more diverse data. Additionally, the consistency of the hand's appearance and of the extracted hand poses can be further improved.

[1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

B. Qualitative Results

Fig 1 shows more examples of the constructed paired training data. We train all the models with a uniform mixture of inpainted and SDEdited object images.

Fig 2 shows that the proposed paired data construction is robust and can be applied to the EPIC-KITCHEN dataset.

Fig 3 shows more comparisons of the generated HOI images by the proposed method (LDM-version as reported in tables) and other image synthesis baselines [3, 5, 8] on the HOI4D dataset.

Fig 4 shows more comparisons of the generated HOI images by the proposed method (LDM-version as reported in tables) and other image synthesis baselines [3,5,8] on EPIC-KITCHEN dataset.

Fig 5 shows more comparisons of the extracted 3D hand pose obtained by the proposed method and other 3D affordance baselines [2, 8] on the HOI4D dataset.

Fig 6 shows more comparisons of the extracted 3D hand pose obtained by the proposed method and other 3D affordance baselines [2, 8] on the EPIC-KITCHEN dataset.

Fig 7 shows an ablation study on comparison of the LDM and GLIDE version of our model on HOI4D and EPIC-KITCHEN datasets.

Fig 8 shows more layout editing results.

Fig 9 shows more results of heatmap-guided synthesis.

References

[1] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. *arXiv*, 2015. 2

[2] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020. 2, 3, 9

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2, 3, 7, 8

[4] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *NeurIPS*, 2015. 1

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014. 2, 3, 7, 8

[6] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *ICML*, 2022. 1, 2

[7] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv*, 2022. 1

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 7, 8, 9

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[10] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 1

[11] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

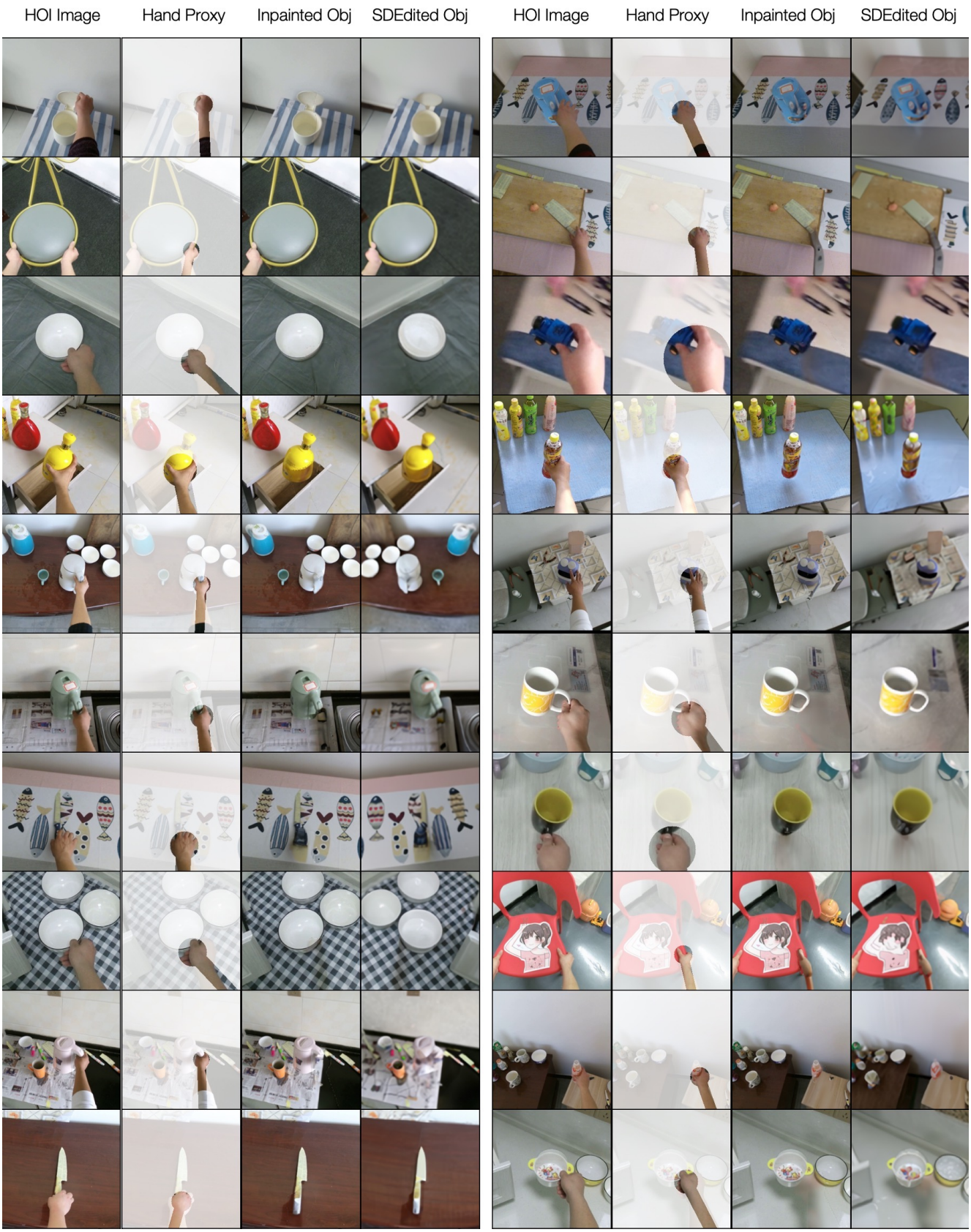

Figure 1. Visualizing more examples of the constructed paired training data. We train all the models with a mixture of inpainted and SDEdited object images.

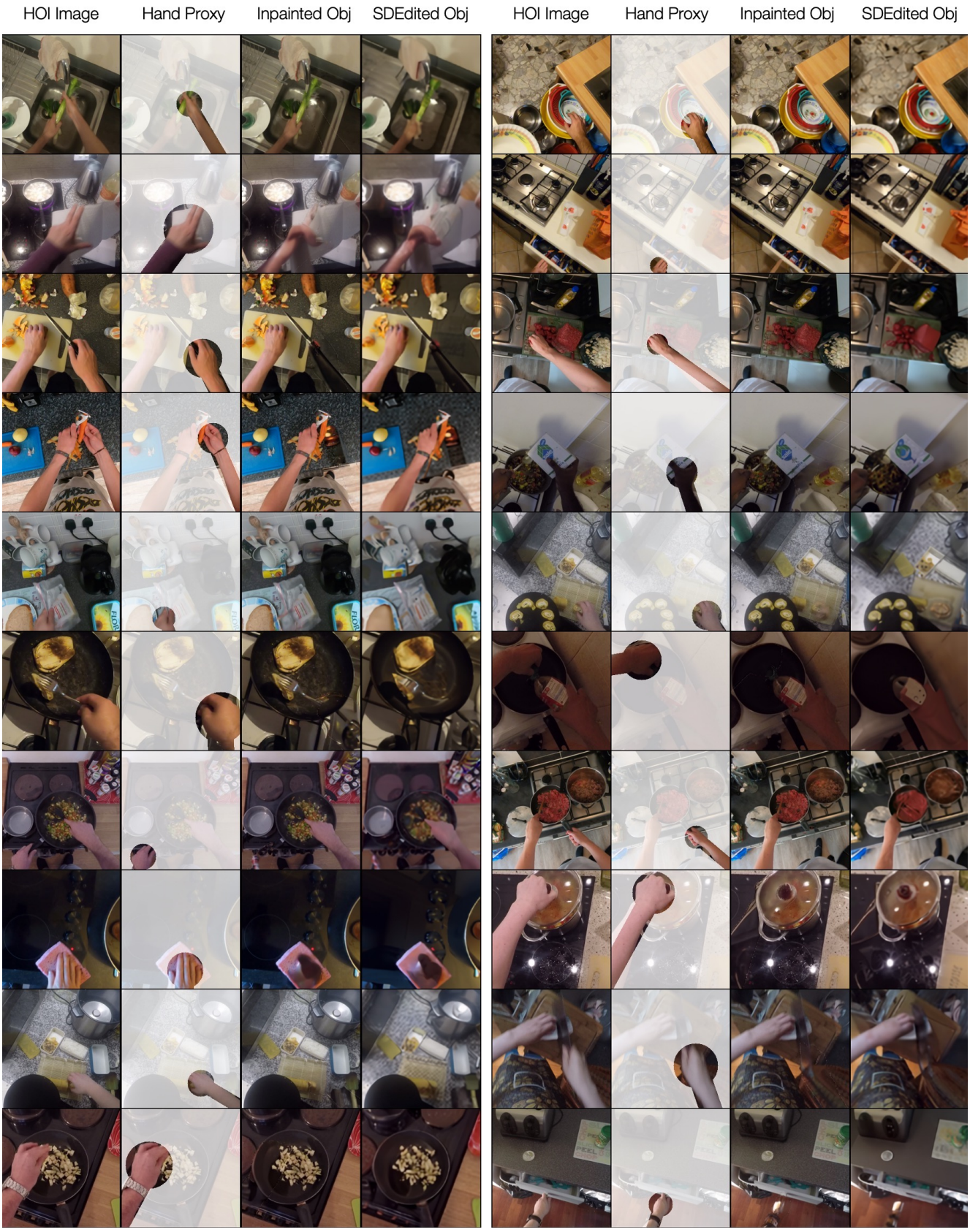

Figure 2. Visualizing the proposed paired data construction applied to EPIC-KITCHEN.

Figure 3. Visualizing more comparisons of the generated HOI images from the proposed method and other image synthesis baselines [3,5,8] on the HOI4D dataset.

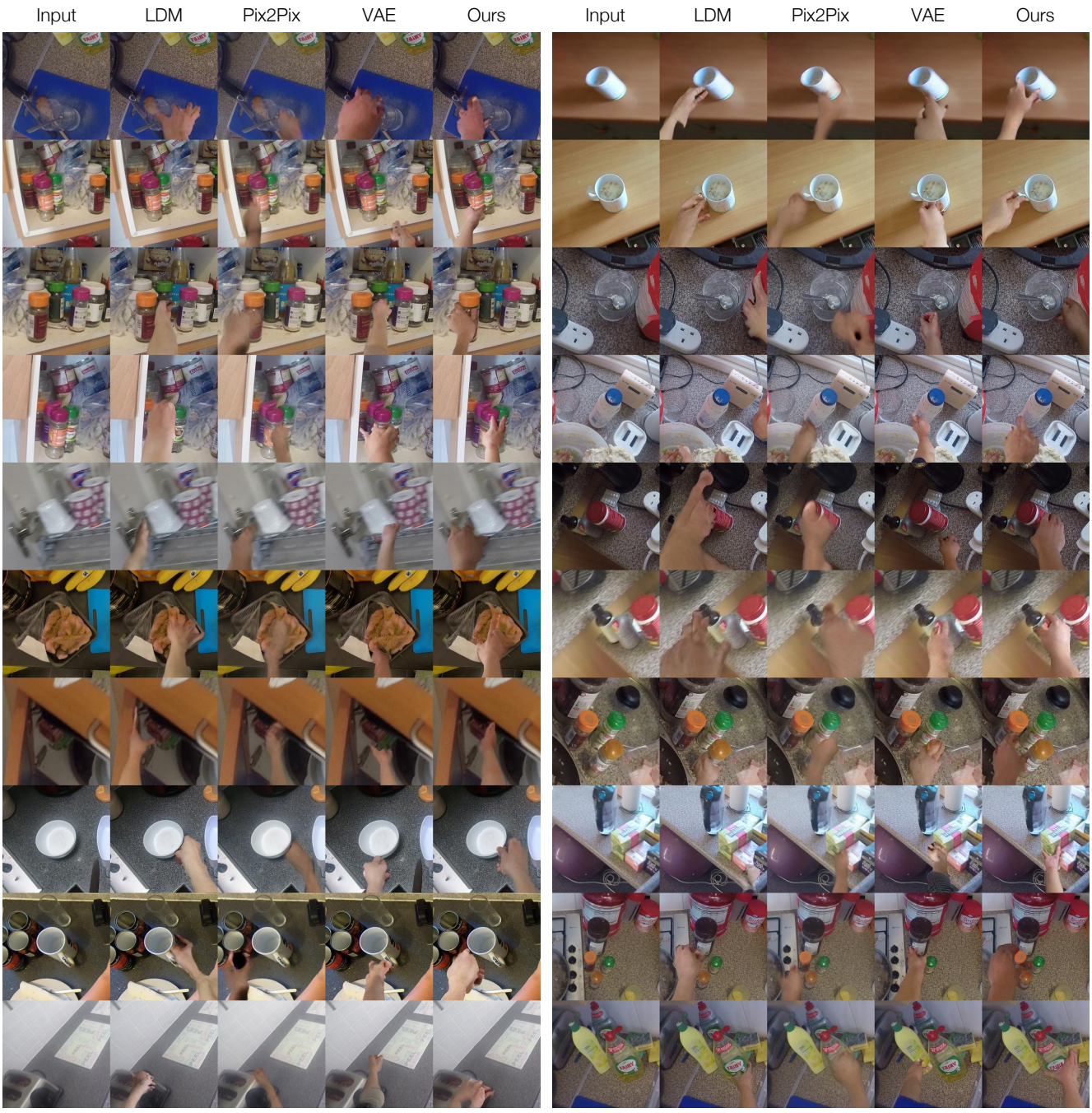

Figure 4. Visualizing more comparisons of the generated HOI images from the proposed method and other image synthesis baselines [3,5,8] on the EPIC-KITCHEN dataset.

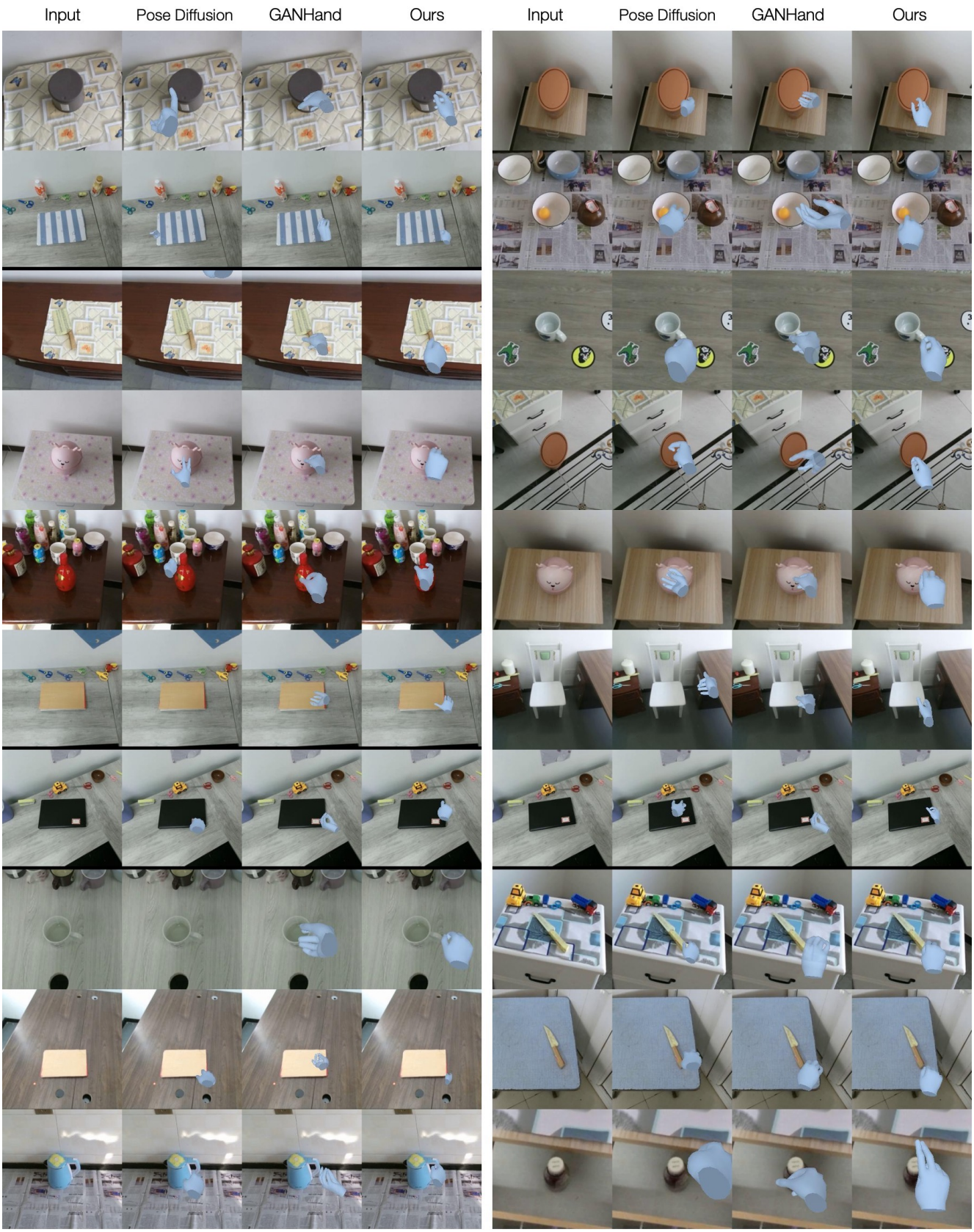

Figure 5. Visualizing more comparisons of the extracted 3D hand pose from the proposed method and other 3D affordance baselines [2, 8] on the HOI4D dataset.

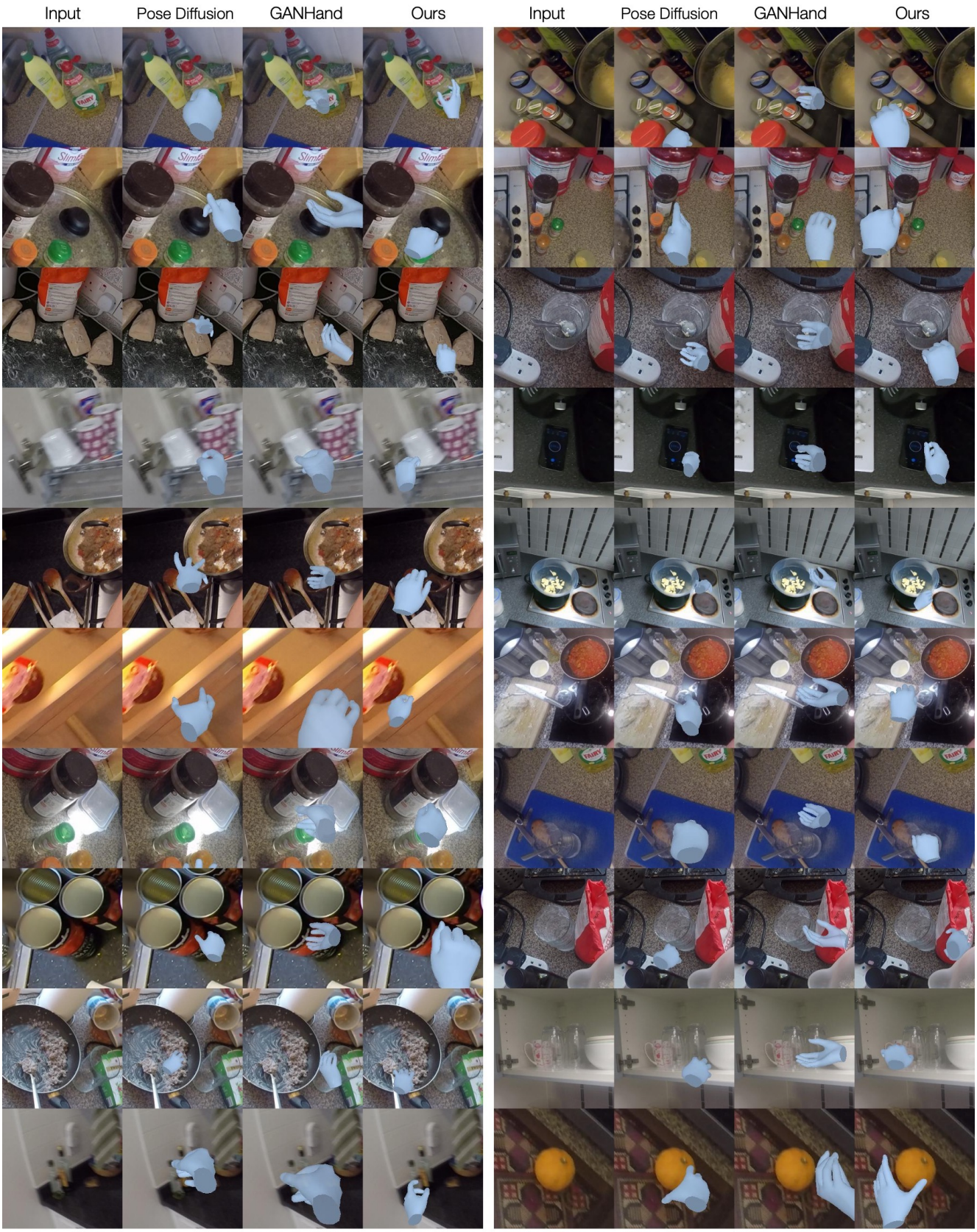

Figure 6. Visualizing more comparisons of the extracted 3D hand pose from the proposed method and other 3D affordance baselines on the EPIC-KITCHEN dataset.

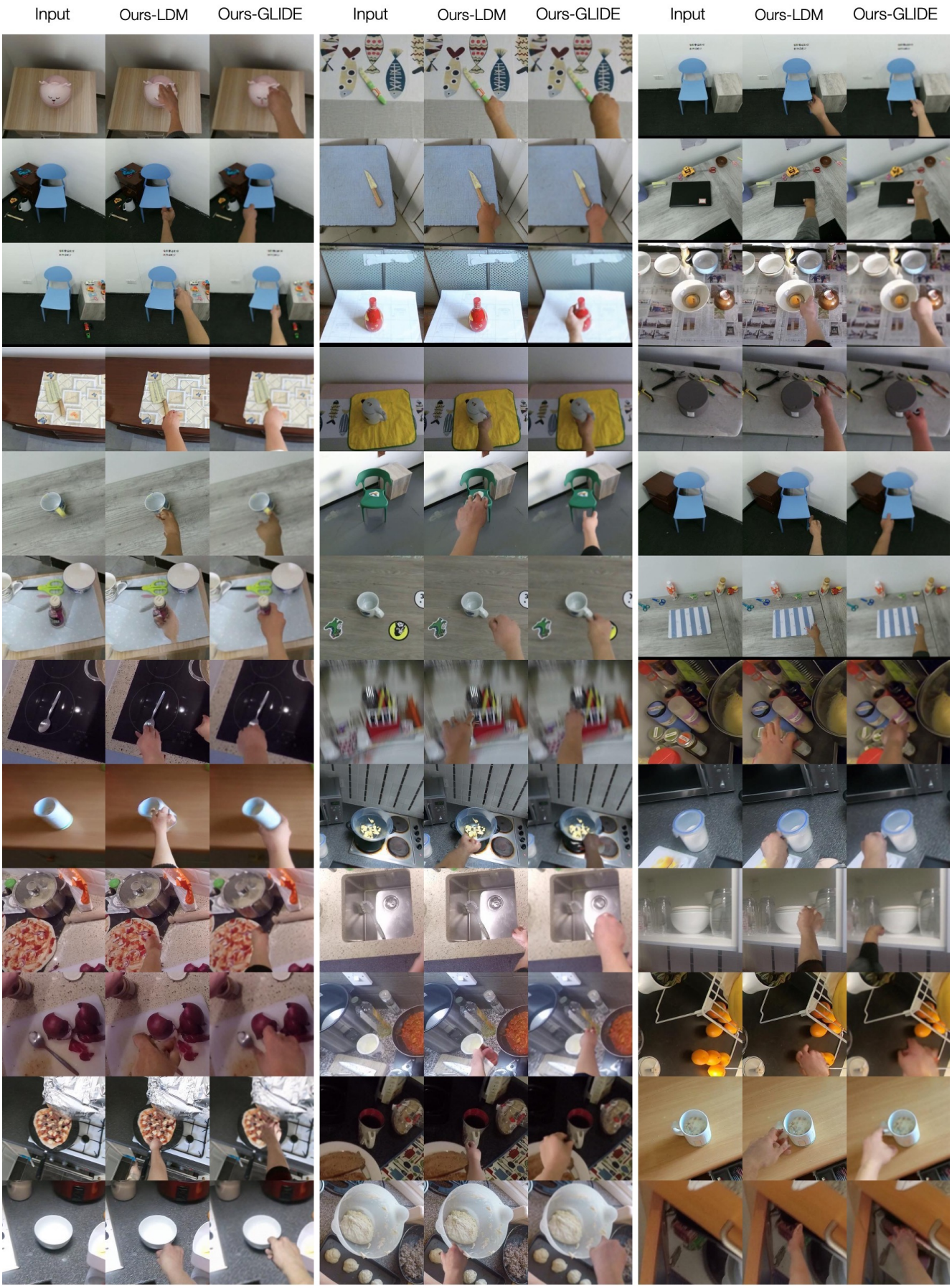

Figure 7. Visualizing the ablation of ContentNet for its LDM-based and GLIDE-based implementations (Sec A.2).

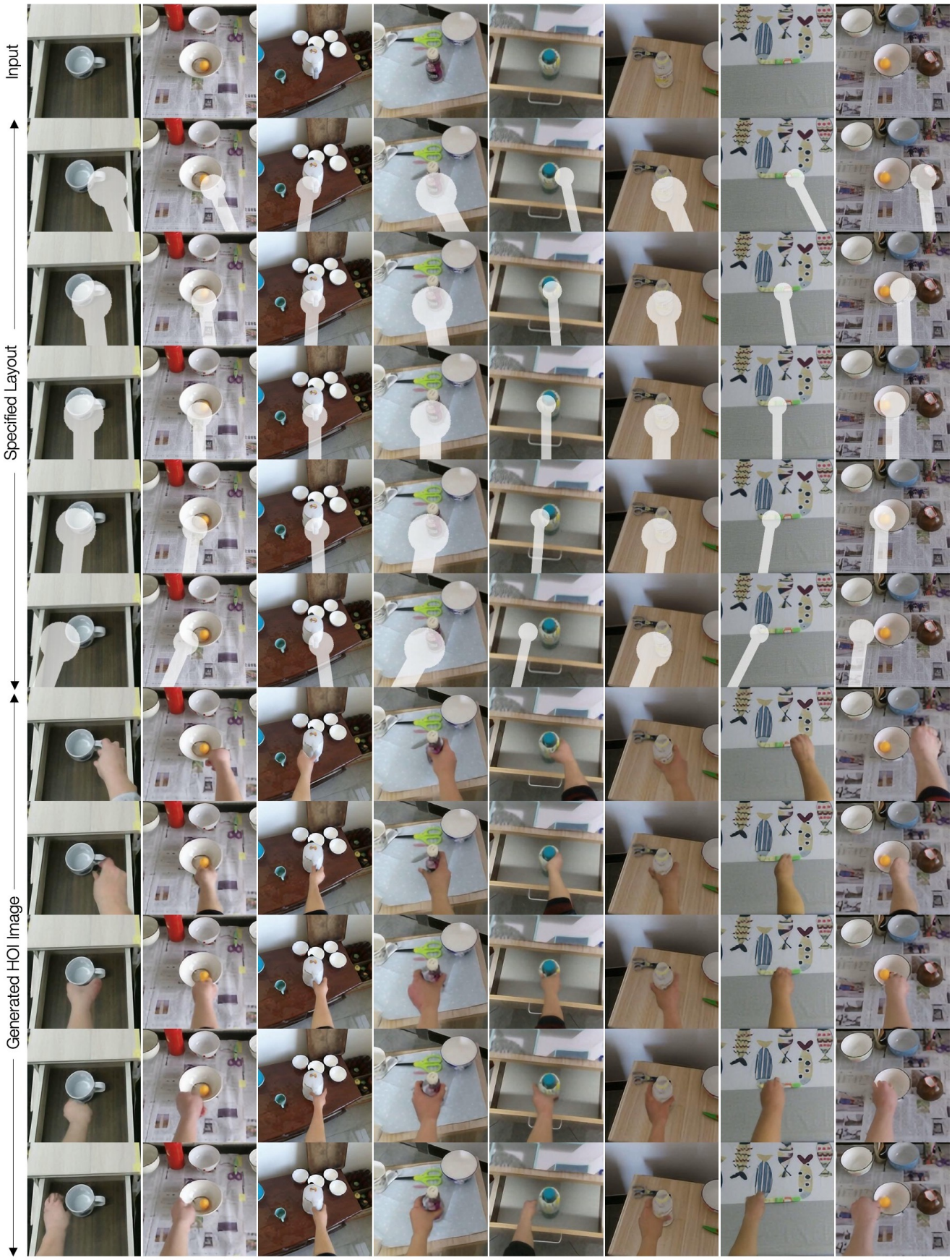

Figure 8. Visualizing more layout editing results.

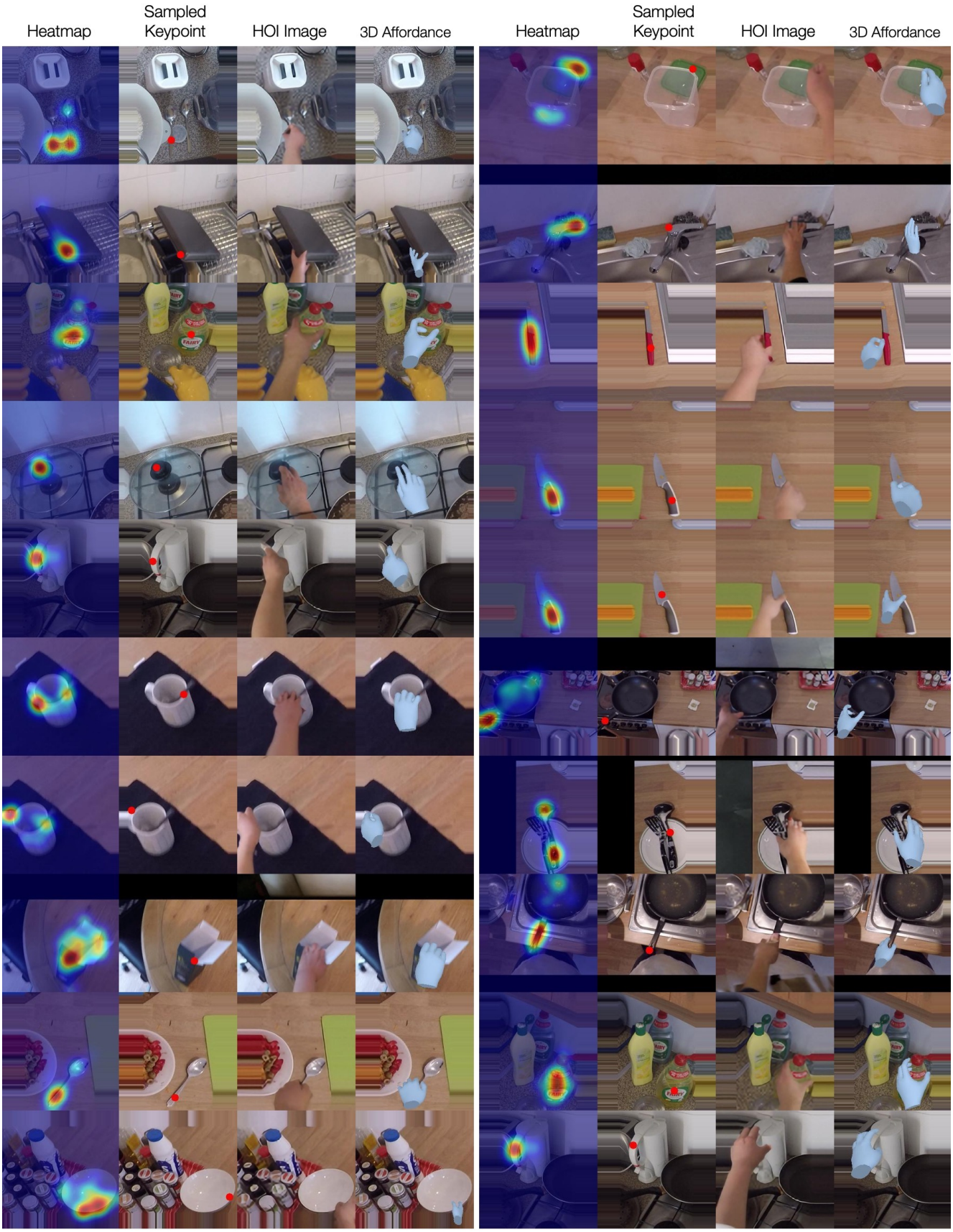

Figure 9. Visualizing more results of heatmap-guided synthesis.

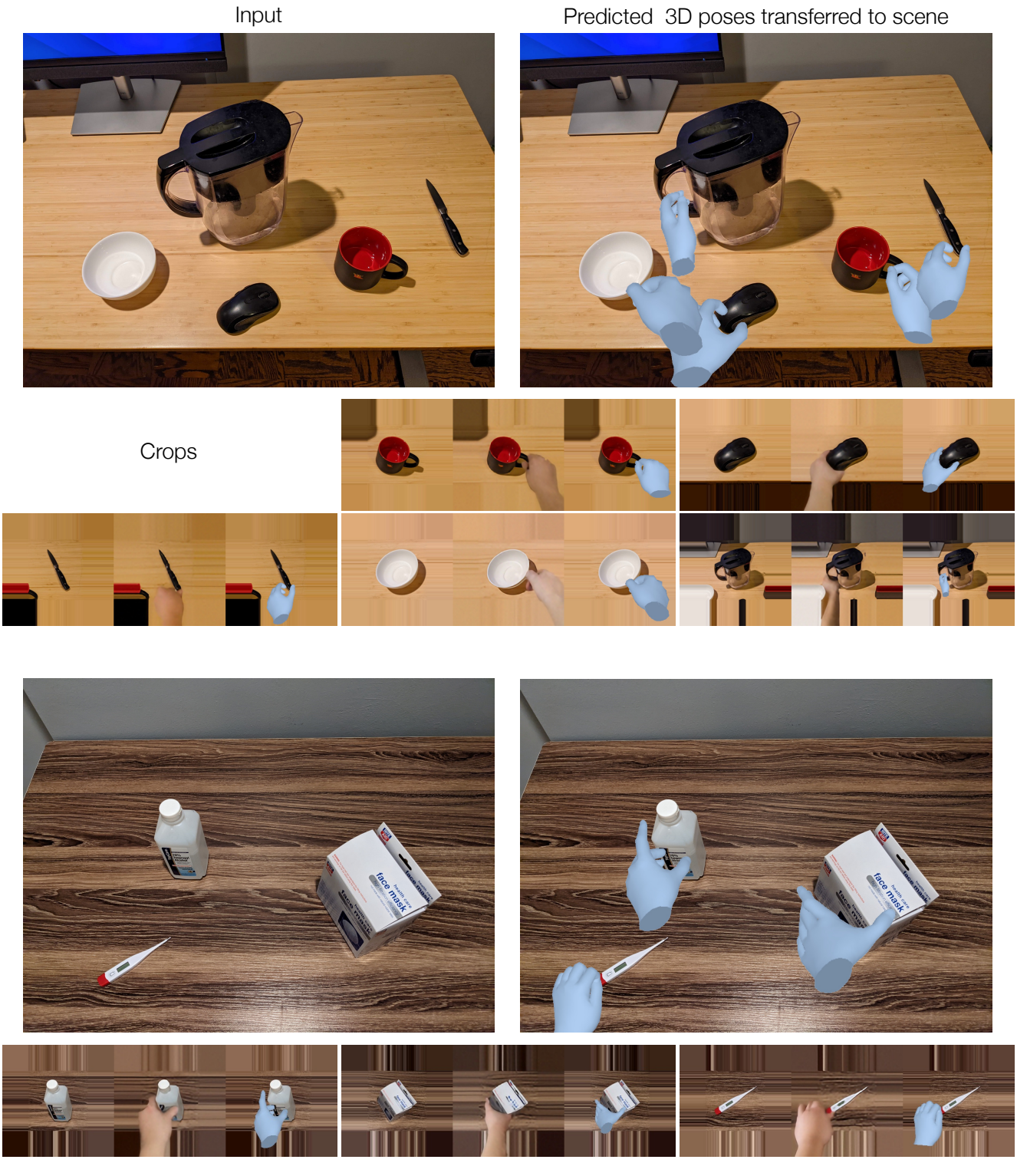

Figure 10. Visualizing more scene integration results with the individual prediction from crops.